



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Antonio Orru'  
25/09/2021



# Outline

---

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

# Executive Summary

---

- Summary of methodologies:
  - Data Collection: API, SQL and Web Scraping
  - Data Wrangling
  - EDA with Matplotlib, Seaborn and SQL
  - Interactive maps with Folium and dashboards with Plotly Dash
  - Machine Learning Classification with Sklearn
- Summary of all results
  - Data Analysis Result
  - Interactive Visualization
  - Classification Prediction

# Introduction

---

- Project background and context:

In the new world of rocket launches for commercial uses one of the cost voices that allow the company to save a lot is the reusability of the first stage: in fact looking at the SpaceX websites, Falcon 9 launch cost 62 million dollars meanwhile the other companies' cost is upward of 165 million. By using SpaceX's data we want to advantage our company SpaceY leaded by our Ceo Allon Mask.

- Problems you want to find answers:

- What are those variables that can lead the first stage to a successful landing?
- How much can each of them affect the landing?
- Which is the ideal mix of this parameters?



Section 1

# Methodology

# Methodology

---

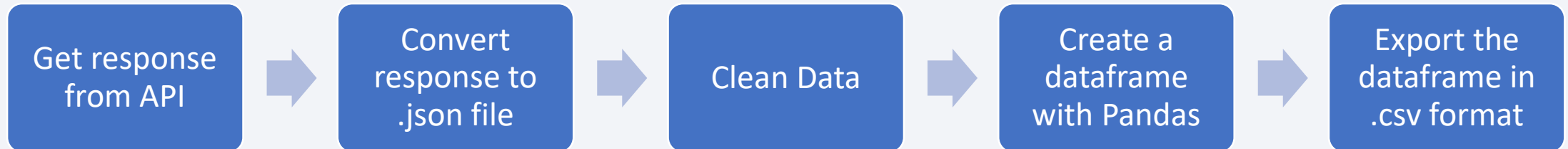
## Executive Summary

- Data collection methodology:
  - SpaceX Rest Api
  - Web Scraping
- Perform data wrangling
  - Fix missing values
  - Feature Selection
  - One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

# Data Collection – SpaceX API

---

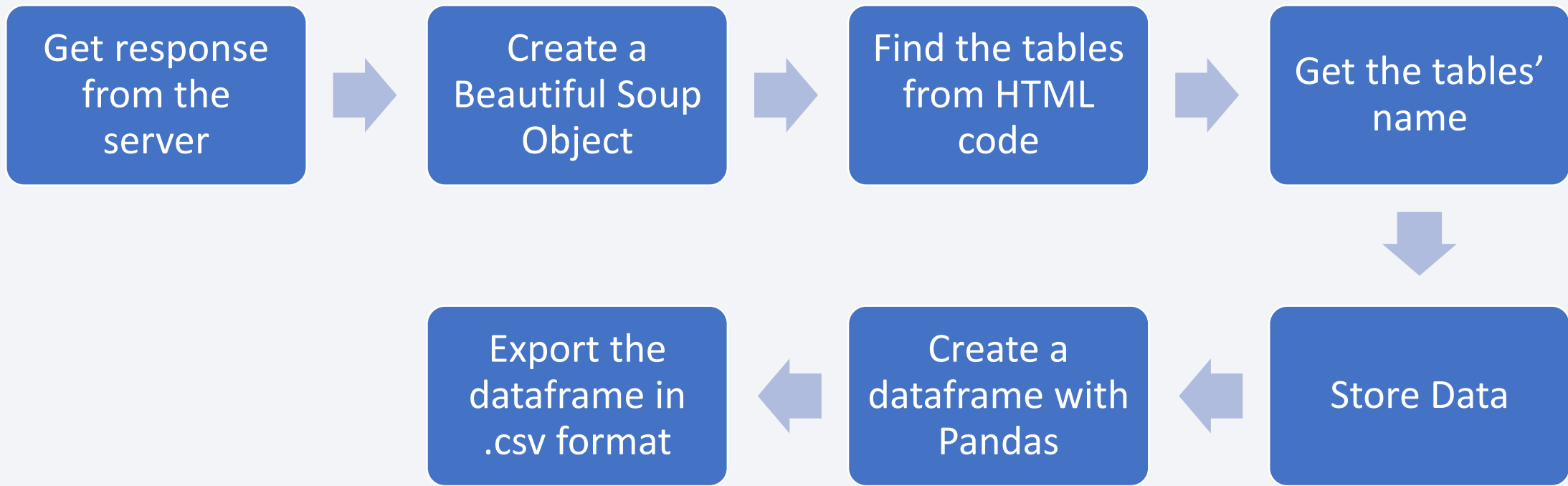
- Data collection is the first part of the job of the data scientist, it consists in cumulate the data that will be studied. The data can come from many sources and techniques, like in this case from web scraping and Rest API.
- This is the process that I followed to collect data from Rest APIs ([Link to GitHub](#)):



# Data Collection - Scraping

---

- This is the process that I followed to collect data using Web Scraping ([Link to GitHub](#)) :

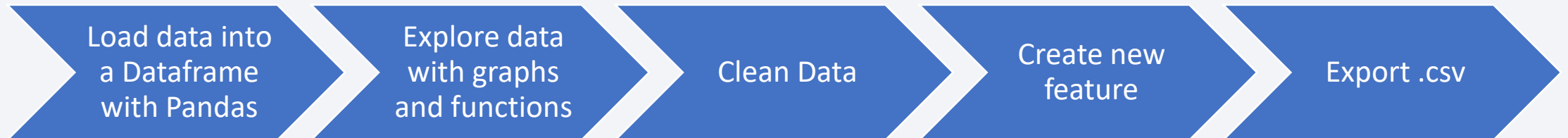




# Data Wrangling

---

- After collecting data, they must be refined through the process of Data Wrangling, this is because they need to be ready to feed the machine learning algorithm. In fact, Data Wrangling is the process that allow us to clean and elaborate data that are we working to. ([Link to GitHub](#))



# EDA with Data Visualization – Scatter Plots

---

- Scatter plots become useful when the Data Scientist needs to find correlations between features. Once the plot is casted, if a pattern appeared we can have a correlation between two or more features. ([Link to GitHub](#))
- I used this plot to cast:
  - Flight Number and Payload
  - Flight number and Launch Site
  - Payload and Launch Site
  - Flight Number and Orbit Type
  - Payload and Orbit type

# EDA with Data Visualization – Bar Graphs

---

- When it's needed to compare different class quantities, like how many success launch for each orbit type it's handy to use Bar Graphs
- I used this plot to cast:
  - Success rate of each orbit type

# EDA with Data Visualization – Line Plot

---

- Line plots are like scatter plots and is used to visualize continuous numeric data type. From line plot we can see easily the trend between to variables and with this kind of graph I plotted:
  - The launch success yearly trend.

# EDA with SQL

---

- Sql is the acronym of Structured Query Language, it's a standardized programming language used to manage databases, but also helpful to study data by using that to interrogate the database. Since the most of data is stored into database it's one of the powerful tools that Data Scientist and Analyst can rely. ([Link to GitHub](#))
- I used sql to answer question like:
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.



# Build an Interactive Map with Folium

---

- Folium is used to create interactive maps; in this case it became useful for geolocalize the location of the different launch site using the following elements ([Link to Github](#)):
  - Circle - Used to indicate launch sites.
  - Marker – Used to display text such labels with location name or for display simple symbols like a map pointer.
  - MarkerCluster – Used to group markers when the map zoom out.
  - MousePosition – Used to display coordinates when your mouse pointer is on the map.
  - Polyline – Used to connect two different point.
- Thanks to Folium we created a map with:
  - Circles that indicate launch sites
  - Green or Red colored marks for each successful or failed launch.
  - Lines that connect point of interest with a marker that indicate the distance between them.
  - A cursor that display coordinates.

# Build a Dashboard with Plotly Dash

---

- Thanks to Plotly Dash it's possible to create an interactive dashboard with elements like ([link to Github](#)):
  - Dropdown menu used to select different launch site.
  - Pie Chart of landing success for each launch site and if one of the sites is selected it shows fail or success landing.
  - Range slider to select the payload of the booster.
  - Scatter chart of landing outcomes vs payload mass divided by booster version.
- Thank to this dashboard is easy to make data exploration.

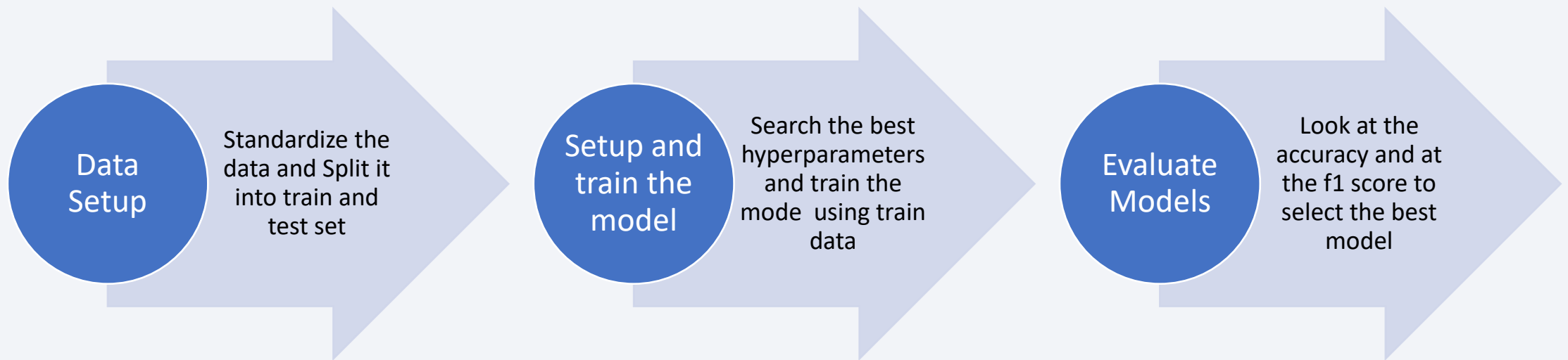
# Predictive Analysis (Classification)

---

- All the refined data that I used in the steps before now are all ready to take the final step: feed the machine learning algorithms. In this research I used classification algorithms :
  - Logistic Regression
  - Decision Tree
  - SVM (Support Vector Machines)
  - KNN (K-nearest neighbors)
- The steps that I followed to train the algorithms and to chose which algorithm should I use are ([Link to GitHub](#)):
  - Standardize the data to help the algorithms to output better results.
  - Split the data into test and train.
  - Using GridSearchCV and train data I tuned and trained my algorithms.
  - Looking at accuracy, confusion matrix,F1 Score and using test data, I decided which algorithm should I use to predict successful launches.

# Predictive Analysis (Classification)

---



# Results

---

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



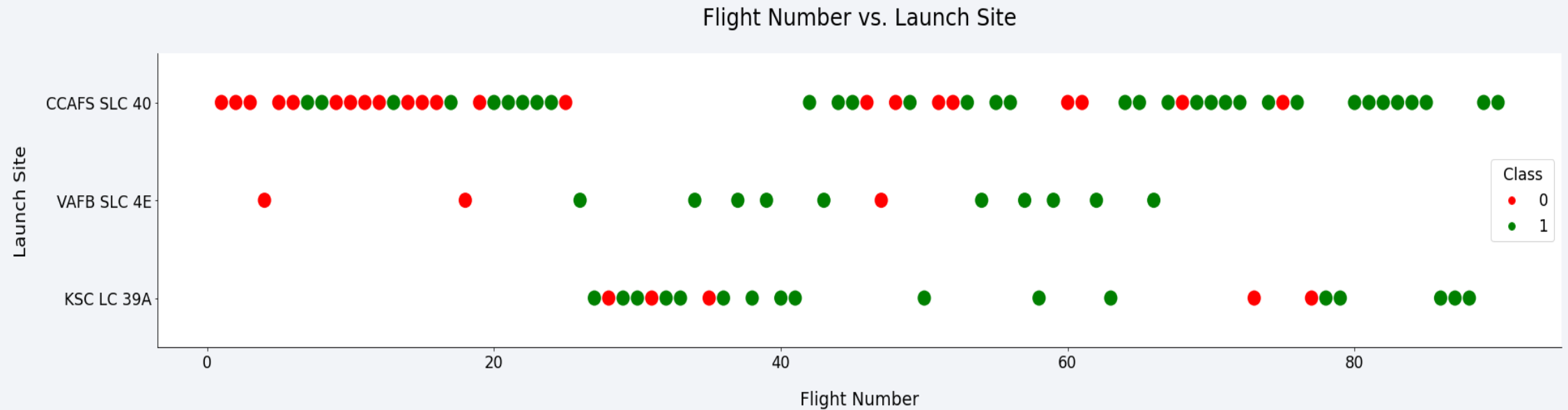
The background of the slide is an abstract composition. It features a solid blue area on the left side, which transitions into a dynamic pattern of diagonal streaks in shades of blue and red on the right. These streaks are layered over a fine, light-colored grid, creating a sense of depth and movement, reminiscent of a digital or data visualization theme.

Section 2

# Insights drawn from EDA

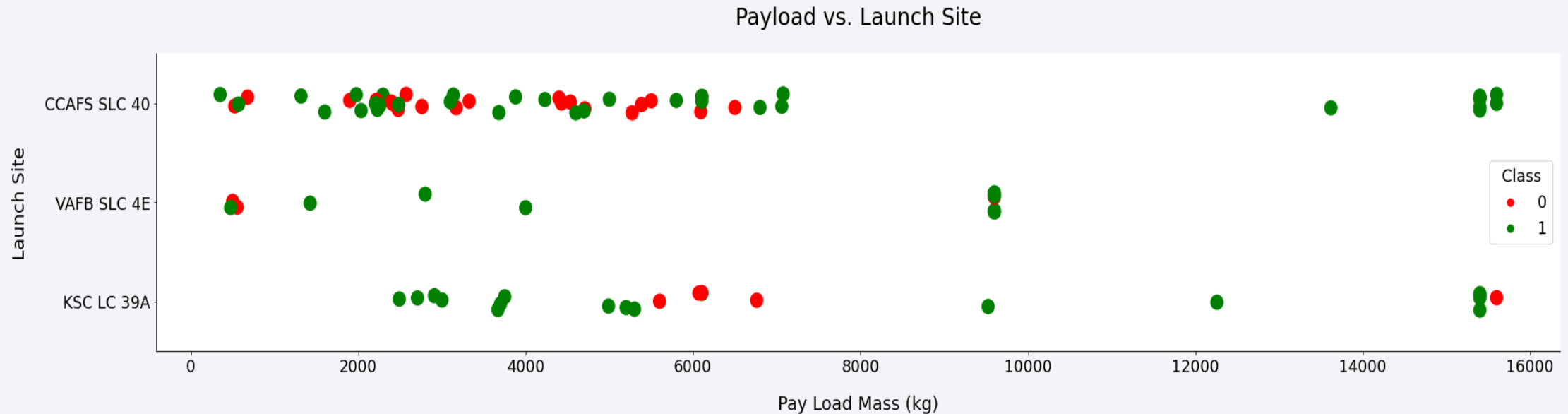


# Flight Number vs. Launch Site



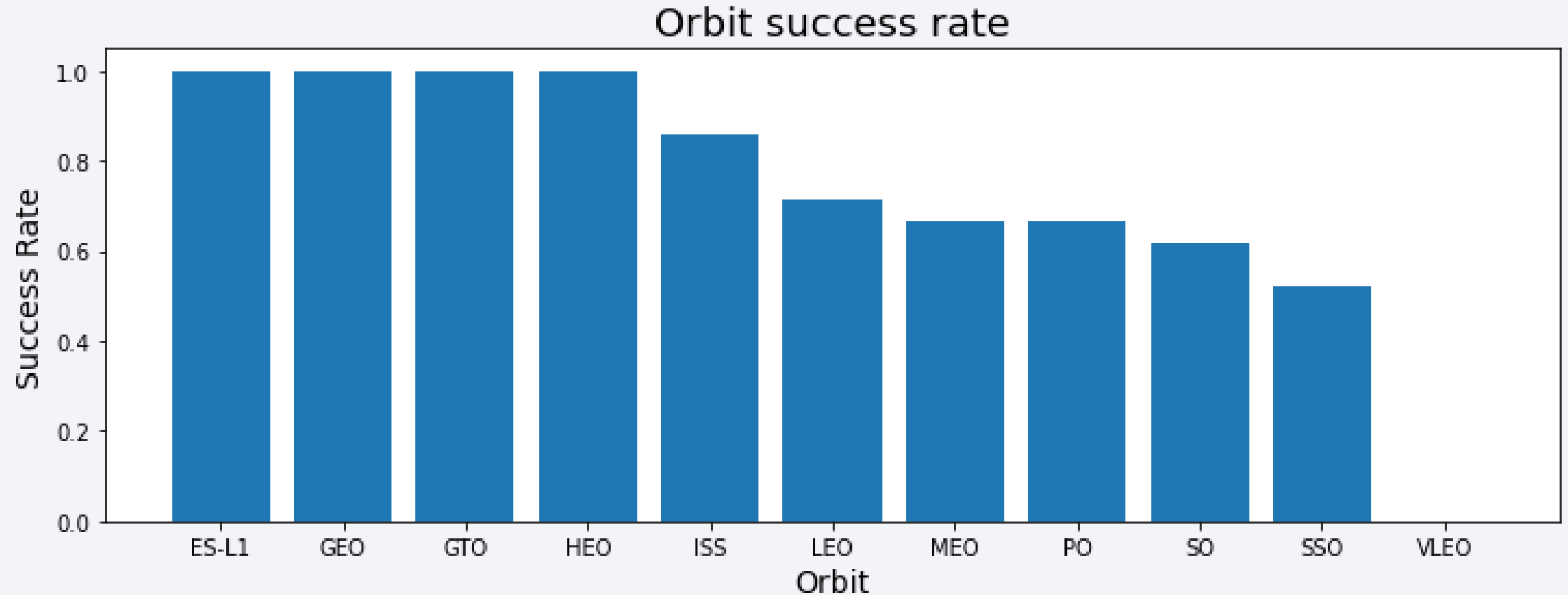
For all of them, starting from the launch number 20, there is an incrementing percentage of successful in launches.

# Payload vs. Launch Site



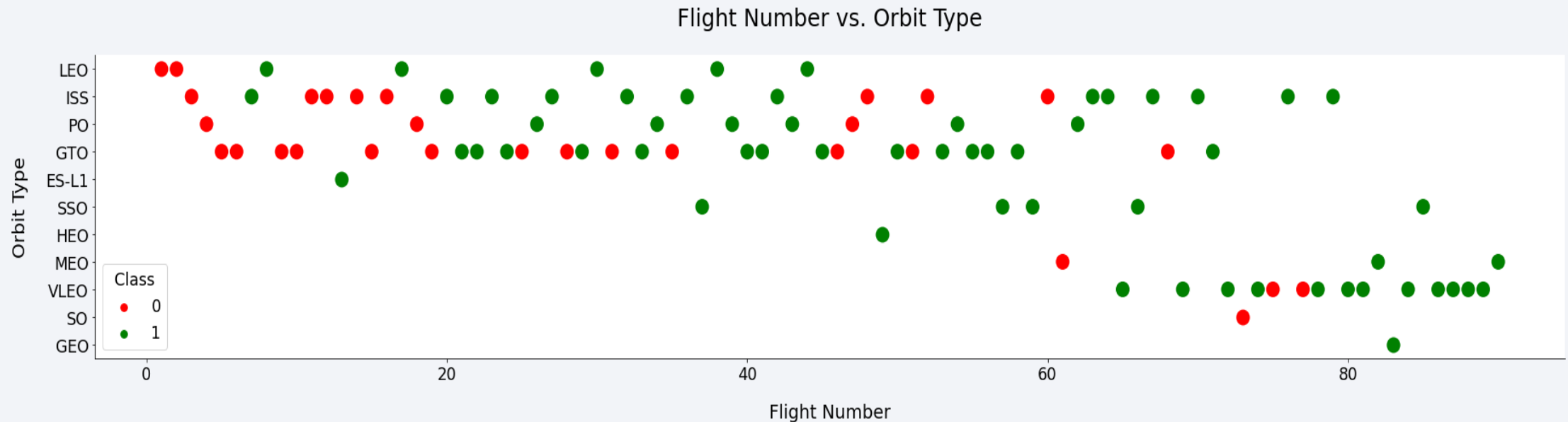
- The high payloads guarantee the best success ratio in all sites.
- The mid-low payloads in CCAFS SLC 40 doesn't guarantee success.
- There is a high success rate in the payloads from about 1000 and 5000 kilos in the VAFB SLC 4E and KSC LC 39A sites

# Success Rate vs. Orbit Type



- If the minimum success rate is set at 80% the best orbits are: ES-L1, GEO, GTO, HEO, ISS

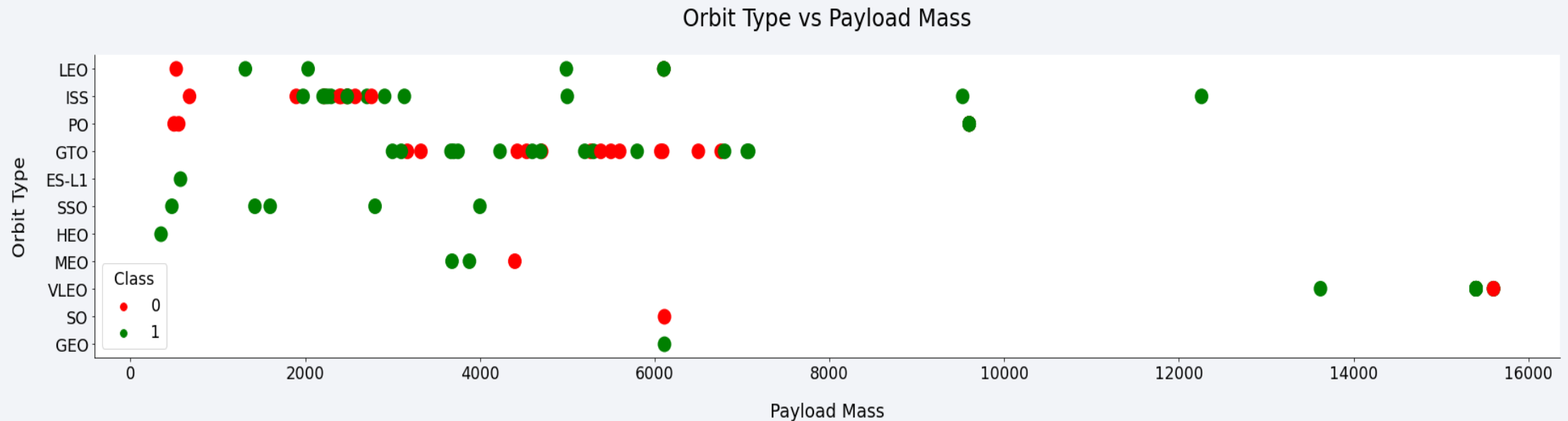
# Flight Number vs. Orbit Type



- First launches went in LEO, ISS, PO, GTO and after the launch number 30 they started to go in the others orbits.



# Payload vs. Orbit Type



- Lower payloads in LEO, ISS and PO means a fail, meanwhile in ES-L1,SSO, HEO means success.
- Higher Payloads in ISS and PO work better than lower Payloads


# Launch Success Yearly Trend

---



- Year after year the success rate grow up, probably because of the experience done launch after launch.
- Only in 2018 and in 2020 the curve fell down for some reason.

# All Launch Site Names

*Display the names of the unique launch sites in the space mission* 

```
%sql select distinct(launch_site) from spacextbl
```

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- Launch site names are: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A and VAFB SLC-4E

# Launch Site Names Begin with 'CCA'

*Display 5 records where launch sites begin with the string 'CCA'*

```
%sql select * from spacextbl where launch_site like 'CCA%' limit 5
```

```
* ibm_db_sa://dqv28731:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

DATE	time__utc__	booster_version	launch_site	payload	payload_mass__kg__	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- We can retrieve only rows about launch site from Cape Canaveral Air Force Station

# Total Payload Mass

---

*Display the total payload mass carried by boosters launched by NASA (CRS)*

```
%sql select sum(payload_mass__kg_) as "total_payload_mass_by_NASA" from spacextbl where customer='NASA (CRS)'
```

```
* ibm_db_sa://dqv28731:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

total_payload_mass_by_nasa
45596

- SpaceX has launched 45596 kilos for NASA



# Average Payload Mass by F9 v1.1

---

*Display average payload mass carried by booster version F9 v1.1*

```
%sql select avg(payload_mass__kg_) as "average_payload_mass_F9" from spacextbl where booster_version like 'F9 v1.1%'
```

```
* ibm_db_sa://dqv28731:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

average_payload_mass_f9
-------------------------

2534
------

- The average payload mass carried by booster version F9 v1.1 is 2534 kilos

# First Successful Ground Landing Date

*List the date when the first successful landing outcome in ground pad was achieved.*

*Hint: Use min function*

```
%%sql select min(DATE) as "first_landing_ground_pad"
from spacextbl
where mission_outcome='Success'
and landing__outcome='Success (ground pad)'
```

```
* ibm_db_sa://dqv28731:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

first_landing_ground_pad
--------------------------

2015-12-22
------------

- The first successful landing outcome in ground pad achieved was in 22-12-2015

# Successful Drone Ship Landing with Payload between 4000 and 6000

*List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000*

```
%%sql select booster_version
from spacextbl
where landing__outcome='Success (drone ship)'
and payload_mass__kg_
between 4000 and 6000
```

```
* ibm_db_sa://dqv28731:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

**booster\_version**

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

- The names of the boosters which have success in drone shipping and have payload mass greater than 4000 but less than 6000 are: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2,

# Total Number of Successful and Failure Mission Outcomes

*List the total number of successful and failure mission outcomes*

```
%sql select mission_outcome ,count(mission_outcome) as count from spacextbl group by mission_outcome
```

```
* ibm_db_sa://dqv28731:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- The total number of successful and failure mission outcomes are: 1 Failure, 99 Success and 1 Success with the payload status unclear

# Boosters Carried Maximum Payload

*List the names of the booster\_versions which have carried the maximum payload mass. Use a subquery*

```
%sql select distinct(booster_version) from spacextbl where payload_mass__kg_=(select max(payload_mass__kg_) from spacextbl)
* ibm_db_sa://dqv28731:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od81cg.databases.appdomain.cloud:31198/bludb
Done.
```

booster_version
-----------------

F9 B5 B1048.4
---------------

F9 B5 B1048.5
---------------

F9 B5 B1049.4
---------------

F9 B5 B1049.5
---------------

F9 B5 B1049.7
---------------

F9 B5 B1051.3
---------------

F9 B5 B1051.4
---------------

F9 B5 B1051.6
---------------

F9 B5 B1056.4
---------------

F9 B5 B1058.3
---------------

F9 B5 B1060.2
---------------

F9 B5 B1060.3
---------------

- 12 different booster version have carried the maximum payload mass

# 2015 Launch Records

*List the failed landing\_outcomes in drone ship, their booster versions, and launch site names for in year 2015*

```
%%sql
select date,landing__outcome,booster_version,launch_site
from spacextbl
where landing__outcome='Failure (drone ship)'
and DATE like '2015%'
```

```
* ibm_db_sa://dqv28731:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.
Done.
```

DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- There were only two failed drone ship landing in 2015



# Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

*Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order* ¶

```
%%sql
select landing__outcome,count(landing__outcome) as count
from spacextbl
where DATE between '2010-06-04'
and '2017-03-20' group by landing__outcome order by count desc
```

```
* ibm_db_sa://dqv28731:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- The most frequent landing outcome was no attempt, the second and the third one are both landing in a drone ship, 5 are Failure and 5 are Success

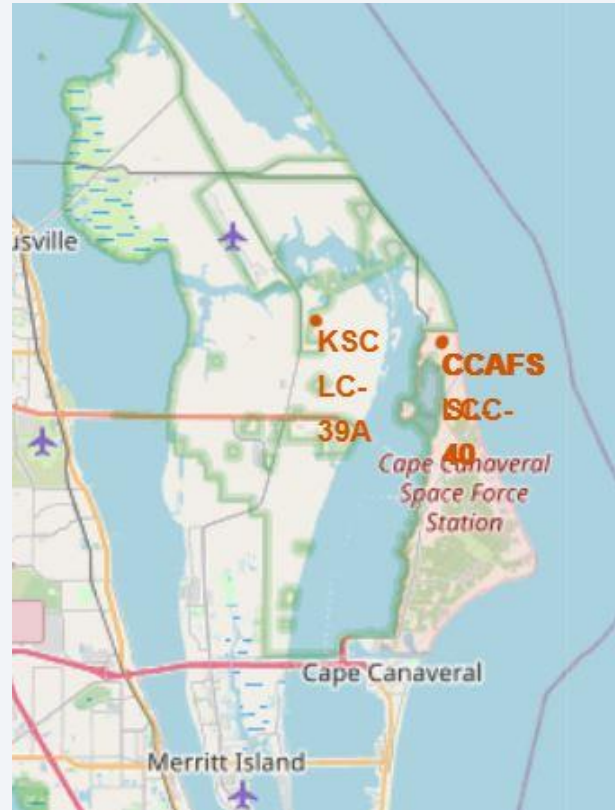
A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a dark blue sky with stars and a view of the Earth's surface, which is illuminated by city lights. The text is overlaid on the left side of the image.

Section 4

# Launch Sites Proximities Analysis

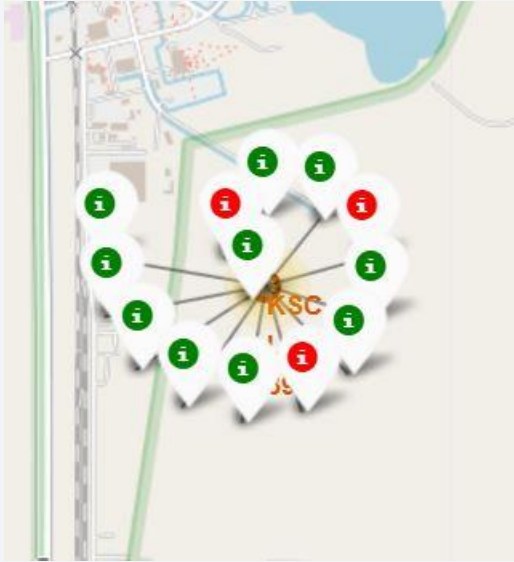


# Launch Site Locations

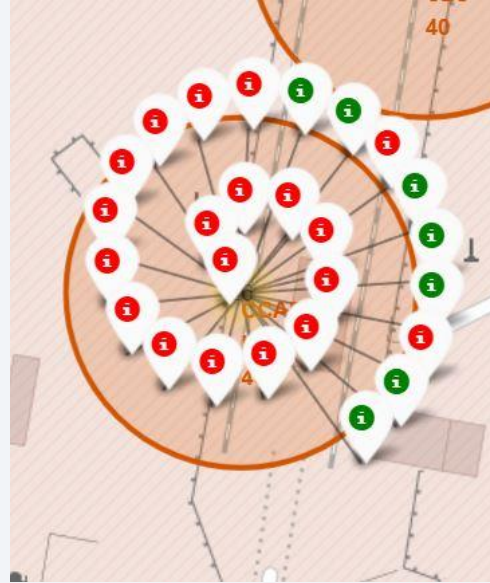


- There are a total of 4 sites: 1 on the east side in California and 3 on the west side in Cape Canaveral, Florida

# Landing Outcome Visualization



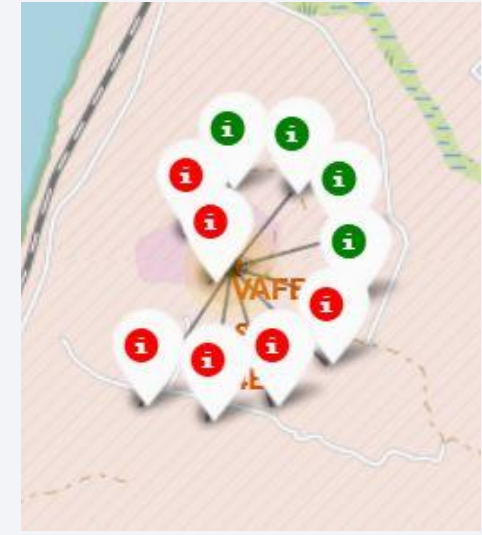
KSC LC-39A



CCAFS LC-40



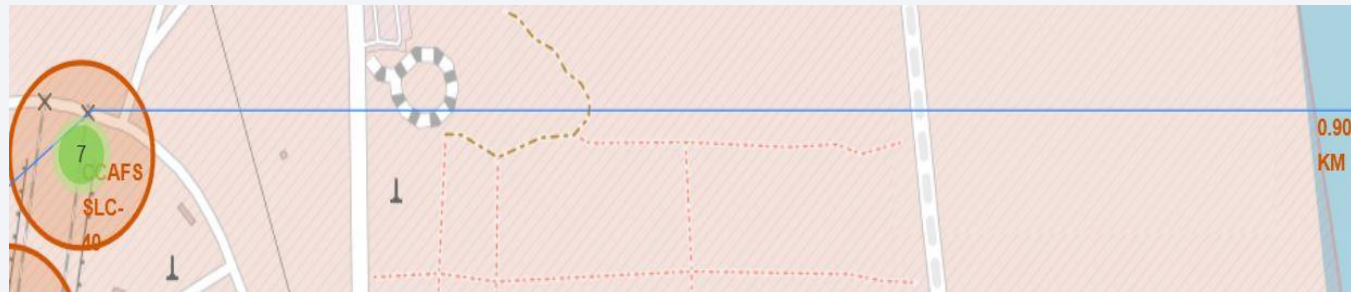
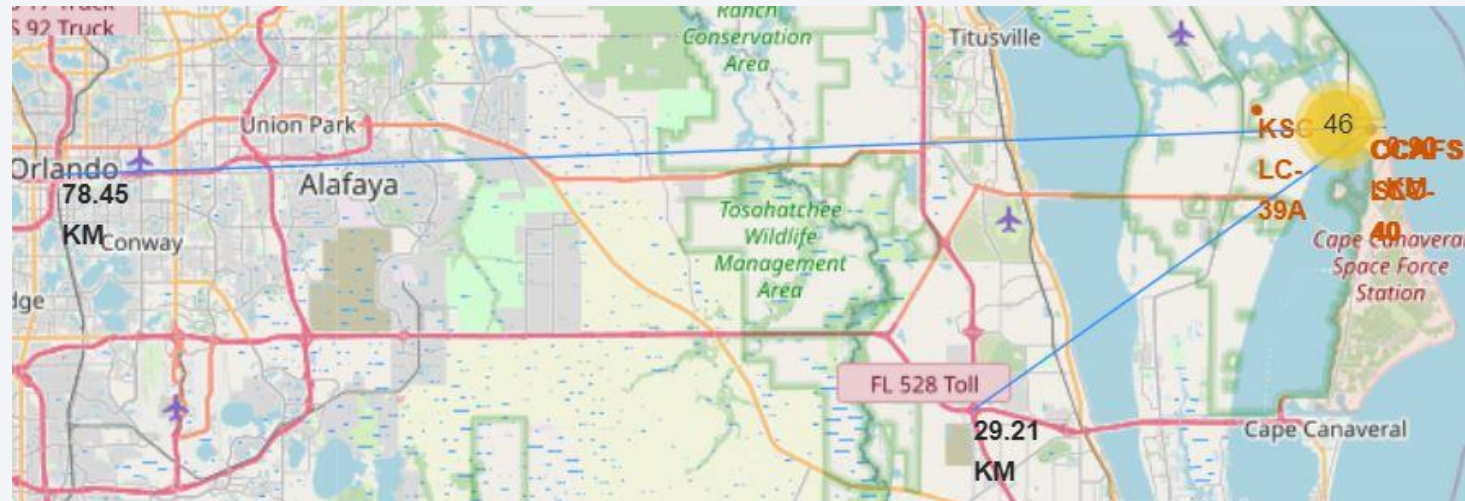
CCAFS SLC-40



VAFB SLC-4E

- Each launch have is own marker, if it's green is a successful booster recovery and if it's red is an unsuccessful.
- We can find the higher booster recovery in KSC LC-39A and the lowest is in CCAFS LC-40

# Launch Site Proximities



CCAFS SLC-40: 0.9 km from the coastline, 29.21 km from the highway and 78.45 km from Orlando

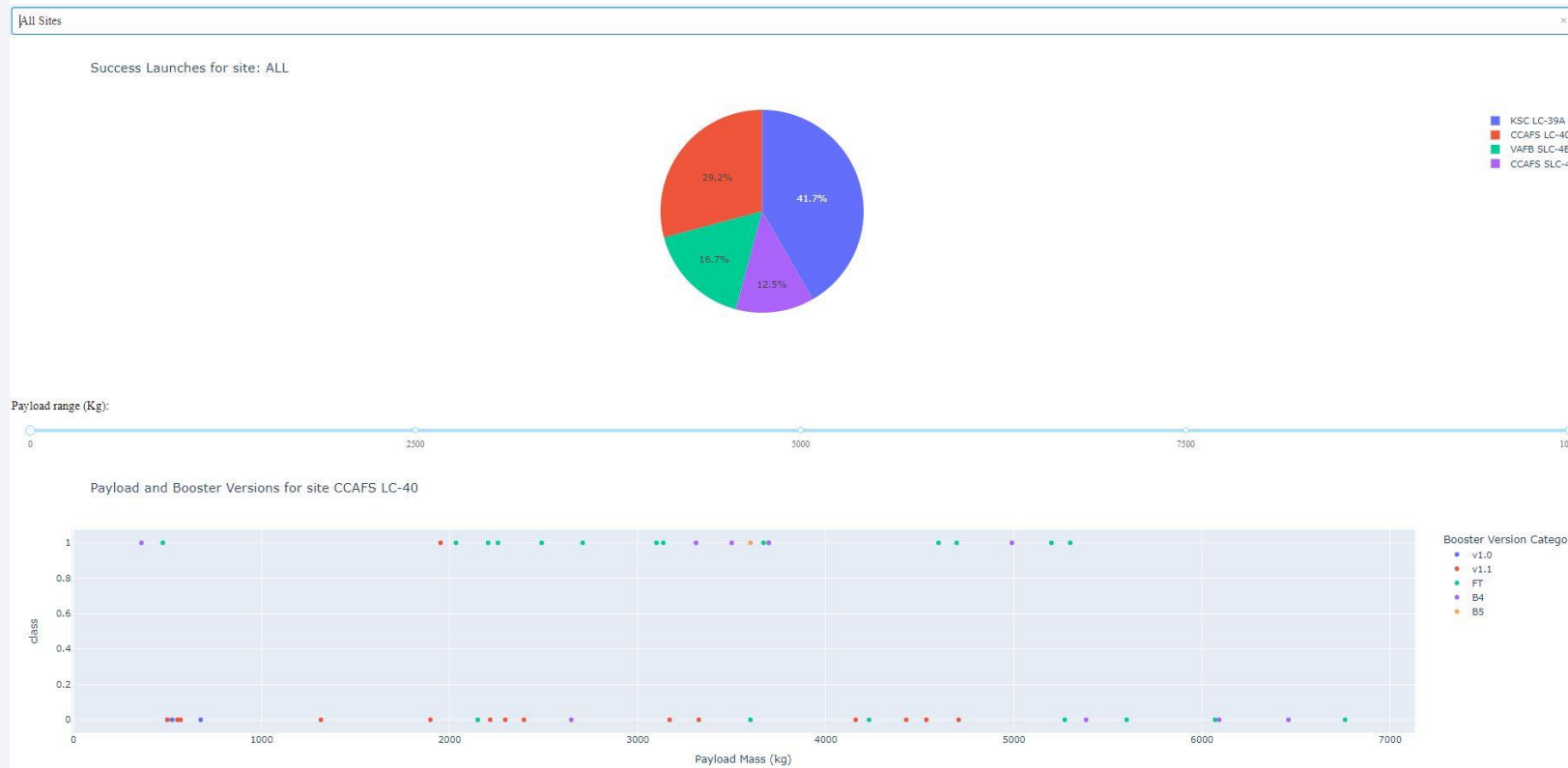




Section 5

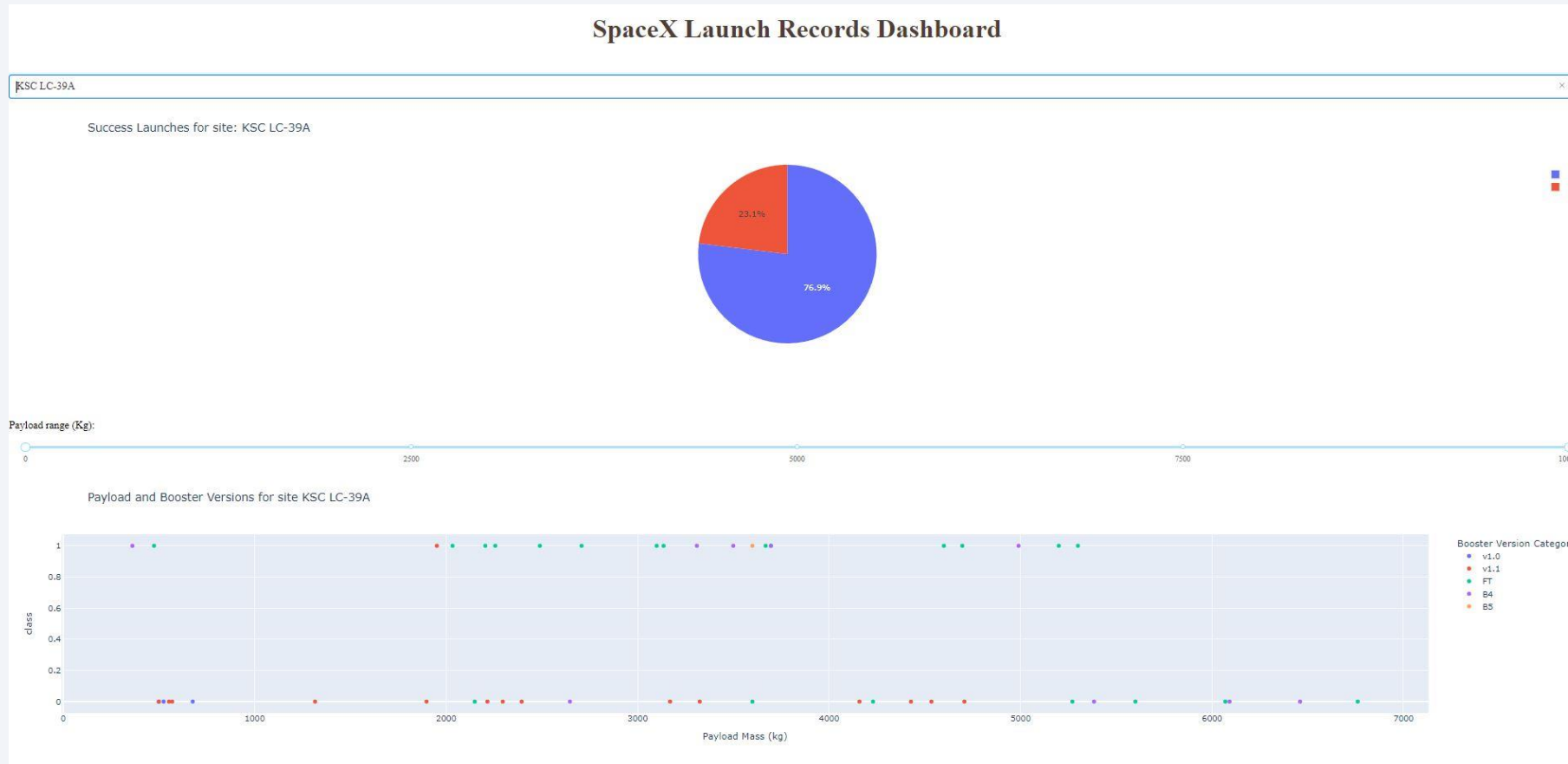
# Build a Dashboard with Plotly Dash

# Dashboard – All Launch Sites



- KSC LC-39A have the highest number of booster recoveries with 41.7%
- CCAFS SLC-40 have the lowest number of booster recoveries with 12.5%

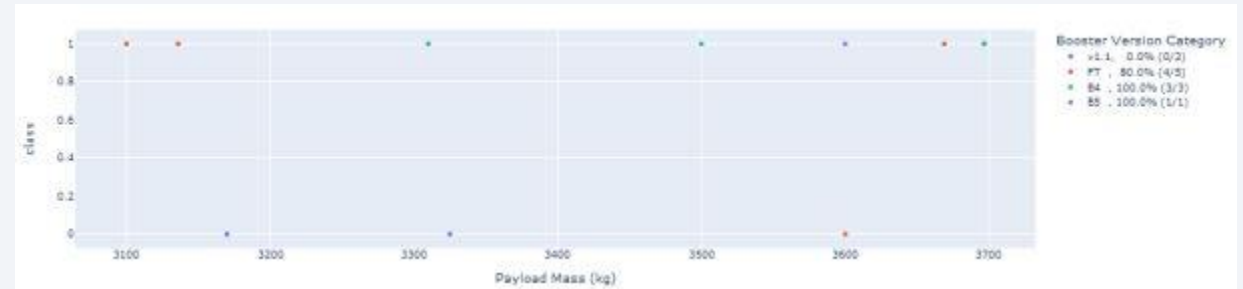
# Dashboard – Most successful launch site



KSC LC-39A has the most successful recovery launch site and it has a success rate of 76.9%

# Dashboard – Payload vs. Launch Outcome

- The successful range of payload mass is between 3000 and 4000 kilos with 7 success over 10 launches.
- The worst range of payload mass is between 6000 and 9000 kilos with 0 success over 4 launches.





Section 6

# Predictive Analysis (Classification)

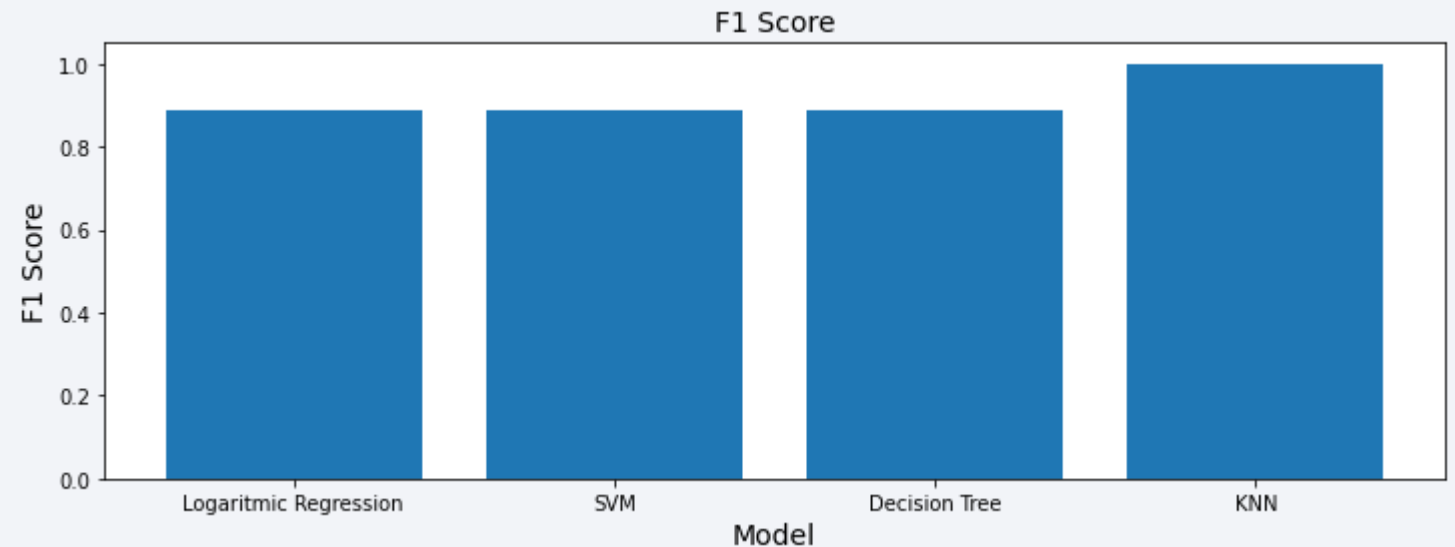


# Classification Accuracy

---

F1 score is a measure of test accuracy and is calculated from the precision and recall of the test.

The test that have performed better is K-nearest neighbors so it's the chosen one.



# Confusion Matrix

- This is the confusion matrix of KNN that I chose. The model predicted 6 times over 6 True Negative and 12 times over 12 True Positive. In conclusion I can say that is the best model based on the test set.



# Conclusions

---

- All launch sites are far away from their neighboring cities, near coastlines and relatively far from highways.
- If the minimum success rate is set at 80% the best orbits are: ES-L1, GEO, GTO, HEO, ISS
- Lower payloads perform better than higher payloads
- Launch site KSC LC-39A is the best for this kind of mission
- Year after year the recovery of the first stage become better
- The best prediction model is the K-nearest neighbors.

Thank you!

